San Diego Housing Midterm Test

Notes

a. Note: investment properties have a positive rental income every month
   i. Rent + hoa > mortgage price
b. Note: investment properties are the ones that increase the most in value over time
c. You are given two datasets, you can use both , amalgamate or not, and you can scrape additional data.e.g., for latent variables

Questions to answer in your narrative

1. Where should an investor, invest? What houses would you recommend they buy as an investment property?
   a. Find the golden cluster for investment : lowest cost house that has the greatest probability of price increase in the next few years.

2. Predict the estimated average selling price in the market using given data
3. Using given listed houses, show the market trends, Using this market trends highlight the low , medium and high values houses.
4. Based on the data derive the following:
   a. Most buyer interested houses
   b. Average buyer interested houses
   c. Less buyer interested houses
5. Scrape the location related information / latent variables [choose 3 out of 5 factors/latent variables]
   a. Proximity to restaurants
   b. Proximity to main highways
   c. Proximity to shopping
   d. Walkability index
   e. Lowest crime area
6. If you buy a house today how much will the price change(profit or loss, increase or decrease) after 2 years of your purchase?

7. Good luck!

Rubric:

1. business case and value-- what hypotheses are you trying to prove?
2. data narrative
3. visualizations, of data prep using first data enrichment (add dataset to base data set)
4. feature importance; gini score
5. feature transformation ; transform features add to dataset, compare results with original
6. second data enrichment -- get an amalgamation; each enrichment enables you to implement / use more algorithms as needed.
7. Third data enrichment -- scrape data from a source and amalgamate
8. implement ml algorithms to build models
    1. Prepare, train and Apply algorithms :you can use the muller loop
        1. **cluster**: GMM, K-means,
        2. **classify**: LogReg, SVM, XGBoost,
        3. **Regression**: LinReg, Ran Forest, KNN,
        4. **dim reduction**: PCA
        5. **probabilistic models**: NaiveBayes.
        6. **deep nets**: MLP
    2. Compare relevant tasks in the same table.
        1. cluster
        2. classify
        3. regress
        4. dim reduc
    3. Write a data narrative to interpret results of each algorithm
9. Suggest Latent Variables or Latent Manifolds, add then to the features and see how prediction results change
10. use metrics for measuring models :
    1. confusion matrix, probab of each slot (e.g., true positives, false positives etc)
    2. Compare in a table:
        1. assess accuracy, precision, recall, f1, rmse
        2. variance, bias,
        3. Probability distributions of

update data narrative with conclusions, comparisons in table(s)

Please submit : on canvas:

1. link to colab,
2. download the notebook and submit the .ipynb ,
3. link to your data set(s) on shared drive
4. pickle and load models